Claims

What is claimed is:

1. A processing system comprising:

a dedicated collective offload engine coupled to a switch fabric of a distributed computing environment having multiple processing nodes also coupled to the switch fabric; and

wherein the dedicated collective offload engine provides collective processing of data from at least some processing nodes of the multiple processing nodes and produces a result based thereon, said result being forwarded to at least one processing node of the multiple processing nodes.

2. The processing system of claim 1, wherein the dedicated collective offload engine is implemented as a hardware device coupled to the switch fabric.

3. The processing system of claim 1, wherein the dedicated collective offload engine comprises:

a payload memory configured to receive and store the data from the at least some processing nodes of the multiple processing nodes; and

an arithmetic logic unit (ALU) coupled to the payload memory, wherein said ALU is configured to retrieve and perform the collective processing of data stored in the payload memory.

4. The processing system of claim 3, wherein the dedicated collective offload engine further comprises:

a dispatcher coupled to the ALU and in communication with the at least some processing nodes of the multiple processing nodes via the switch fabric, said dispatcher configured to control the collective processing of the data from the at least some processing nodes of the multiple processing nodes and the sharing of the result based thereon.

5. The processing system of claim 4, wherein the dedicated collective offload engine further comprises:

at least one task table coupled to the dispatcher, wherein the at least one task table is configured to store task identification information related to the at least some processing nodes of the multiple processing nodes; and

at least one synchronization group table coupled to the dispatcher, wherein the at least one synchronization group table is configured to store identification information related to one or more groups of the at least some processing nodes of the multiple processing nodes.

6. The processing system of claim 4, wherein the dedicated collective offload engine further comprises:

an adapter coupled to the switch fabric, wherein said adapter is configured to communicate with the switch fabric using a link protocol; and

interface logic coupled to the adapter, the payload memory and the dispatcher, wherein the interface logic facilitates communication between said adapter and said payload memory and between said adapter and said dispatcher.

7. The processing system of claim 1, wherein the processing system further comprises a plurality of dedicated collective offload engines in communication with one another via the switch fabric, wherein said communication facilitates the collective processing of data from the at least some processing nodes of the multiple processing nodes and the producing of the result based thereon.

8. The processing system of claim 1, wherein the processing system further comprises a plurality of dedicated collective offload engines in communication with one another via a channel disposed therebetween, said channel being independent of the switch fabric, and wherein said communication facilitates the collective processing of data from the at least some processing nodes of the multiple processing nodes and the producing of the result based thereon.

9. The processing system of claim 1, wherein the collective processing provided by the dedicated collective offload engine includes execution of at least one collective operation for the at least some processing nodes of the multiple processing nodes without using a software tree.

10. The processing system of claim 1, wherein the collective processing provided by the dedicated collective offload engine includes managing at least one distributed lock associated with at least one of a distributed database and a distributed file system.

11. A system for processing, said system comprising:

means for providing, by a dedicated collective offload engine coupled to a switch fabric in a distributed computing environment, collective processing of data from at least some processing nodes of multiple processing nodes of the distributed computing environment;

means for producing, by the dedicated collective offload engine, a result based on said collective processing; and

means for forwarding said result to at least one processing node of the multiple processing nodes.

12. A method of processing comprising:

providing, by a dedicated collective offload engine coupled to a switch fabric in a distributed computing environment, collective processing of data from at least some processing nodes of multiple processing nodes of the distributed computing environment;

producing, by the dedicated collective offload engine, a result based on said collective processing; and

forwarding said result to at least one processing node of the multiple processing nodes.

13. The method of claim 12, wherein the dedicated collective offload engine is implemented as a hardware device coupled to the switch fabric.

14. The method of claim 12, further comprising:

receiving and storing, at a payload memory, the data from the at least some processing nodes of the multiple processing nodes, wherein said payload memory is a component of the dedicated collective offload engine; and

retrieving and performing, at an arithmetic logic unit (ALU), the collective processing of data stored in the payload memory, wherein said ALU is a component of the dedicated collective offload engine and is coupled to the payload memory.

15. The method of claim 14, further comprising:

controlling the collective processing of the data from the at least some processing nodes of the multiple processing nodes, wherein said controlling is performed by a dispatcher of the dedicated collective offload engine coupled to the ALU, and in communication with the at least some processing nodes of the multiple processing nodes via the switch fabric; and

controlling, by the dispatcher, the sharing of the result with the at least one processing node of the multiple processing nodes.

16. The method of claim 15, further comprising:

storing, in at least one task table coupled to the dispatcher, task identification information related to the at least some processing nodes of the multiple processing nodes, wherein said at least one task table is a component of the dedicated collective offload engine; and

storing, in at least one synchronization group table coupled to the dispatcher, identification information related to one or more groups of the at least some processing nodes of the multiple processing nodes, wherein said at least one synchronization group table is a component of the dedicated collective offload engine.

17. The method of claim 15, further comprising:

communicating, via an adapter, across the switch fabric using a link protocol, wherein said adapter is coupled to the switch fabric and is a component of the dedicated collective offload engine; and

facilitating, by interface logic, communication between said adapter and said payload memory and between said adapter and said dispatcher, wherein said interface logic is a component of the dedicated collective offload engine.

18. The method of claim 12, further comprising:

communicating among a plurality of dedicated collective offload engines via the switch fabric, wherein said communicating facilitates the collective processing of data from the at least some processing nodes of the multiple processing nodes and the producing of the result based thereon.

19. The method of claim 12, further comprising:

communicating among a plurality of dedicated collective offload engines via a channel disposed therebetween, said channel being independent of the switch fabric, wherein said communicating facilitates the collective processing of data from the at least some processing nodes of the multiple processing nodes and the producing of the result based thereon.

20. The method of claim 12, wherein said providing collective processing includes executing at least one collective operation for the at least some processing nodes of the multiple processing nodes without using a software tree.

21. At least one program storage device readable by a machine, tangibly embodying at least one program of instructions executable by the machine to perform a method of processing comprising:

providing, by a dedicated collective offload engine coupled to a switch fabric in a distributed computing environment, collective processing of data from at least some processing nodes of multiple processing nodes of the distributed computing environment;

producing, by the dedicated collective offload engine, a result based on said collective processing; and

sharing said result with at least one processing node of the multiple processing nodes.

22. The at least one program storage device of claim 21, wherein the dedicated collective offload engine is implemented as a hardware device coupled to the switch fabric.

23. The at least one program storage device of claim 21, said method further comprising:

receiving and storing, at a payload memory, the data from the at least some processing nodes of the multiple processing nodes, wherein said payload memory is a component of the dedicated collective offload engine; and

retrieving and performing, at an arithmetic logic unit (ALU), the collective processing of data stored in the payload memory, wherein said ALU is a component of the dedicated collective offload engine and is coupled to the payload memory.

24. The at least one program storage device of claim 23, said method further comprising:

controlling the collective processing of the data from the at least some processing nodes of the multiple processing nodes, wherein said controlling is performed by a dispatcher of the dedicated collective offload engine coupled to the ALU, and in communication with the at least some processing nodes of the multiple processing nodes via the switch fabric; and

controlling, by the dispatcher, the sharing of the result with the at least one processing node of the multiple processing nodes.

25. The at least one program storage device of claim 24, said method further comprising:

storing, in at least one task table coupled to the dispatcher, task identification information related to the at least some processing nodes of the multiple processing nodes, wherein said at least one task table is a component of the dedicated collective offload engine; and

storing, in at least one synchronization group table coupled to the dispatcher, identification information related to one or more groups of the at least some processing nodes of the multiple processing nodes, wherein said at least one synchronization group table is a component of the dedicated collective offload engine.

26. The at least one program storage device of claim 24, said method further comprising:

communicating, via an adapter, across the switch fabric using a link protocol, wherein said adapter is coupled to the switch fabric and is a component of the dedicated collective offload engine; and

facilitating, by interface logic, communication between said adapter and said payload memory and between said adapter and said dispatcher, wherein said interface logic is a component of the dedicated collective offload engine.

27. The at least one program storage device of claim 21, said method further comprising:

communicating among a plurality of dedicated collective offload engines via the switch fabric, wherein said communicating facilitates the collective processing of data from the at least some processing nodes of the multiple processing nodes and the producing of the result based thereon.

28. The at least one program storage device of claim 21, said method further comprising:

communicating among a plurality of dedicated collective offload engines via a channel disposed therebetween, said channel being independent of the switch fabric, wherein said communicating facilitates the collective processing of data from the at least some processing nodes of the multiple processing nodes and the producing of the result based thereon.

29. The at least one program storage device of claim 21, wherein said providing collective processing includes executing at least one collective operation for the at least some processing nodes of the multiple processing nodes without using a software tree.

30. A data structure facilitating collective processing, said data structure comprising:

a packet to be sent from a processing node, of multiple processing nodes coupled to a switch fabric in a distributed computing environment, to a dedicated collective offload engine also coupled to the switch fabric, said packet comprising:

a first field including an identifier of a collective operation to be executed by said dedicated collective offload engine; and

a second field including a payload, wherein said payload comprises data from said processing node to be collectively processed by said dedicated collective offload engine based on the collective operation.

* * * * *